

The Effect of Multiple Replies for Natural Language Generation Chatbots

Eason Chen
National Taiwan Normal University
Taipei, Taiwan
eason.tw.chen@gmail.com

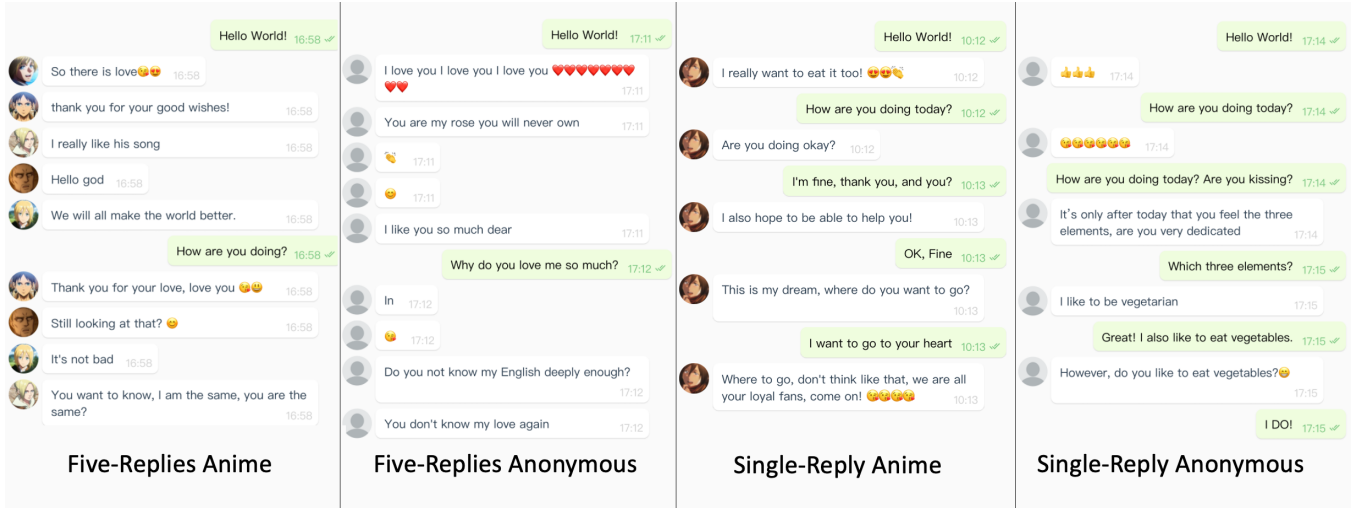


Figure 1: The Screenshot of 2 (single reply vs. five replies) \times 2 (anonymous avatar vs. anime avatar) experimental conditions. From left to right is Five-Replies Anime, Five-Replies Anonymous, Single-Reply Anime, Single-Reply Anonymous, respectively.

ABSTRACT

In this research, by responding to users' utterances with multiple replies to create a group chat atmosphere, we alleviate the problem that Natural Language Generation chatbots might reply with inappropriate content, thus causing a bad user experience. Because according to our findings, users tend to pay attention to appropriate replies and ignore inappropriate replies. We conducted a 2 (single reply vs. five replies) \times 2 (anonymous avatar vs. anime avatar) repeated measures experiment to compare the chatting experience in different conditions. The result shows that users will have a better chatting experience when receiving multiple replies at once from the NLG model compared to the single reply. Furthermore, according to the effect size of our result, to improve the chatting experience for NLG chatbots which is single reply and anonymous avatar, providing five replies will have more benefits than setting an anime avatar.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

CHI '22 Extended Abstracts, April 29-May 5, 2022, New Orleans, LA, USA

© 2022 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-9156-6/22/04.

<https://doi.org/10.1145/3491101.3516511>

CCS CONCEPTS

• Human-centered computing \rightarrow Laboratory experiments.

KEYWORDS

Natural Language Generation, Chatbot, Multiple Replies, Improve Chatting Experience

ACM Reference Format:

Eason Chen. 2022. The Effect of Multiple Replies for Natural Language Generation Chatbots. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '22 Extended Abstracts)*, April 29-May 5, 2022, New Orleans, LA, USA. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3491101.3516511>

1 INTRODUCTION

"Chatbot can be a software which interacts with people using natural language by building a conversation" [3]. Chat-oriented chatbots are becoming popular in providing chatting services [3] thanks to the recently developed neural language modeling [1], especially Natural Language Generation model, like GPT-2 [8].

To generate text, Natural Language Generation (NLG) model adopts a decoding strategy, such as the top-k or Nucleus method [6], to generate one token at a time based on the input text and currently generated tokens until a certain stop criterion is met. Due to the randomness nature of the decoding strategy [6], the generated token sequence will be different even when the same

input sequence is given. Therefore, it is hard to predict the quality of a single output based on the input. Consequently, during the conversation between the user and the chatbot, especially on open-domain topics, the current NLG chatbots inevitably make errors [4] and affect the chatting experience.

We notice that most NLG chatbot researches focus on providing a single reply which is the best output by the best model [3, 9, 11]. However, since the output from the NLG model can't be perfect, we consider another approach to let users maintain a satisfactory chatting experience: providing more replies. In this research, we want to know how multiple replies influence the chatting experience for NLG Chatbot. Therefore, we recruited 102 participants to join a 2 (single reply vs. five replies) \times 2 (anonymous avatar vs. anime avatar) repeated-measures experiment and analyzed with two-way ANOVA to compare the chatting experience in different conditions (see Figure 1).

In sum, the main contributions for this paper are:

- (1) According to our survey, we find users tend to pay attention to appropriate replies and ignore inappropriate replies when chatting with the chatbot.
- (2) Compared to a chatbot with single reply, the result from our experiment shows that users have a better chatting experience when the chatbot responds to users' one utterance with multiple replies.
- (3) According to the effect size of our result, to improve the chatting experience for an NLG chatbot which is single reply and anonymous avatar, providing multiple replies will have more benefits than setting an anime avatar.

2 THEORETICAL FRAMEWORK AND HYPOTHESES

When using a chat-oriented chatbot, the key to having a good chatting experience is to avoid dialog breakdown. Dialog breakdown means users are feeling difficult to continue their conversation [4]. The main reason causing dialog breakdown is inappropriate messages, which includes ignoring the user's input, providing a not understandable reply, repetition, or contradictory content [4]. Therefore, we suppose the frequency of suitable and unsuitable messages can predict the chatting experience. *H1: The frequency of appropriate replies can significantly predict the chatting experience score.*

Because the NLG model's output is highly randomized [6], it is hard to maintain the user's chatting experience. Once users receive an unacceptable answer, they might stop chatting. However, we can view it differently: users will continue chatting if they receive an acceptable reply. So, how to increase the likelihood that users receive a proper response? We can increase the number of replies. In purely rational analysis, this is a question of probability: assuming that given the user's input x , the probability that the model generates a suitable response will be $p(x)$ (ex: 0.4). Then the probability that the model can't generate a suitable response will be $1 - p(x)$ (ex: 0.6). However, if the model generates five parallel responses at once, the probability that the model can't generate any fitting response will be $(1 - p(x))^5$ (ex: $0.65 = 0.0776$). Compared with the single reply, the chance that the user will not receive any proper reply in the multi reply condition will be greatly reduced (ex: from 0.6



Figure 2: The Screenshot of the experimental platform at five-replies anime condition. The left is the chatroom, and the right is the instruction of the test.

to 0.0776). In other words, with more replies, the probability that the model will generate at least one suitable answer will be greatly increased (ex: from 0.4 to 0.9224). Thus, users might have a better chatting experience in the multi replies condition than the single reply condition. This leads to the following hypothesis: *H2: Multiple replies condition has a better user experience than the single reply condition.*

The hypothesis mentioned above be valid or not highly dependent on users' mental processes. According to the Sensemaking theory [10], people will interpret the information they receive in a way that they can understand. "Explicit efforts at sensemaking tend to occur when the current state of the world is perceived to be different from the expected state of the world" [10]. Hence, users will be trying to make sense of what the chatbot wants to say when interacting. Accordingly, suppose we present more replies to create a group chat atmosphere. Then, it might be easier for users to comprehend what the chatbot is saying and increase tolerance for unfitting responses. For example, users might not take those unfitting messages seriously since they assume the chatbot may not be talking to them but to other chatbots in the same chatroom. Therefore, we propose the following hypothesis: *H3: Users tend to pay attention to appropriate replies and ignore inappropriate replies when interacting with chatbots.*

Nonetheless, when providing multiple replies, we received feedback from pre-study participants who said that it is hard to distinguish the replies as 'many from one person' or 'many from many people' without avatar. Therefore, we provide avatars for each reply as a control condition to help participants differentiate the multiple replies. We suppose that at the chatting experience score, there is an interaction effect between the avatar type and the number of replies. *H4: There is an interaction effect between avatar type and the number of replies at chatting experience score.*

3 RESEARCH METHOD

The experiment used a 2 (single reply vs. five replies) \times 2 (anonymous avatar vs. anime avatar) repeated measures design. Each participant was required to undergo four randomly ordered experimental conditions. Apart from that, we choose five replies here

Table 1: Questions to measure the frequency of suitable replies.

Nr.	Statement
1	In the previous chat, there are many ‘appropriate’ replies
2	In the previous chat, there are many ‘inappropriate’ replies

because it is best to fit the chatroom window without overflow while maximizing the replies. And the anime avatars for chatbots are selected from the popular anime “Attack on Titan” (see in Figure 1 and 2).

3.1 Participants and Procedure

We recruited 102 participants in Taiwan from the internet through convenience sampling with a mean age of 22.3 years (range 19 – 50, $SD = 3.89$). At the experiment, participants will first login to the platform with their own computer and read the instruction of the experiment. Following the instruction, participants will have ten times of chat warmup to get familiar with the platform. After that, participants will go through one of the four conditions. Each condition participant will first chat ten rounds (Figure 2) then fill the form for the chat experience of the round. After all four rounds are finished, the participant will fill the final form for suitable response frequency and their attention tendency.

3.2 Measures

All questionnaires are measured by Google Form.

3.2.1 Chatting Experience. We translated The Chatbot Usability Questionnaire (CUQ) [5] into Chinese to measure the chatting experience in each round. The scale is designed to measure the user experience of using chatbots. The participant must select 1 (strongly disagree) to 5 (strongly agree) for sixteen statements (such as The chatbot understood me well, or Chatbot responses were irrelevant). According to the author [5], the score of CUQ has a significant positive correlation with the score from System Usability Scale (SUS) [2] and the User Experience Questionnaire (UEQ) [7]. Furthermore, based on the data we collected, the value for Cronbach’s Alpha for the CUQ was $\alpha = .94$.

3.2.2 The frequency of appropriate replies. To measure the frequency of appropriate replies, we use a custom five-point scale (Table 1). Participants must select 1 (strongly disagree) to 5 (strongly agree) for the following two statements. The frequency of suitable replies will be obtained by the score of the first question, plus six minus the score of the second question. These statements were verified with three experts from relative fields, and based on the data we collected, the value for Cronbach’s Alpha for this survey was $\alpha = .83$. Participants will review the definition of appropriate and inappropriate replies before measurement begins.

3.2.3 Attention/Ignore tendency for appropriate replies and inappropriate replies. To measure participants’ attention tendency, we use a custom five-point scale (Table 2). The participants must select 1 (strongly disagree) to 5 (strongly agree) for the following four statements. These statements were verified with three experts from

Table 2: Questions to measure attention/ignore tendency for appropriate replies and inappropriate replies.

Nr.	Statement
1	In the previous chat, I often pay attention to the appropriate replies
2	In the previous chat, I often pay attention to inappropriate replies
3	In the previous chat, I often ignore appropriate replies
4	In the previous chat, I often ignore inappropriate replies

relative fields, and based on the data we collected, the value for Cronbach’s Alpha for this survey was $\alpha = .81$. To compare users’ attention tendency, we will use the paired samples t-test for questions 1, 2, and 3, 4, respectively.

3.3 Experimental Platform

The experimental platform is a WEB service with front-end and back-end. The front-end build using Vue.js and Bootstrap. The back-end uses Python FastAPI with MongoDB to process the data. We connect Google Sheet API with the platform so participants can fill the questionnaire at Google Form, and the platform will check their progress via Sheet API to make sure they finish and can go to the next round. The platform’s source code is available at GitHub¹.

3.4 NLG Model

About the NLG model, we use the design from Yang and Tseng [11], which is a GPT-2 and BERT dual-model reply generation system. To be specific, after receiving the input message, GPT-2 will first generate many replies and use BERT to predict replies’ coherence score. Then, the system will sort the coherence score and use top n replies as the output. The system can successfully input various dialogues and adjust emotions to output appropriate responses after learning about the 1.7 million corpus provided by the Japanese NTCIR Chinese Emotional Conversation Generation (CECG) evaluation task in 2019 [12]. To control the variance in the experiment, we only use ‘like’ as the emotion type and generate one response per request. Moreover, even though the model is Chinese-based, we connect the Google Translate API so that the chat system can reply according to the detected language. That is why some texts in screenshots are English.

4 RESULT

The data analysis was performed using SPSS 23.0.

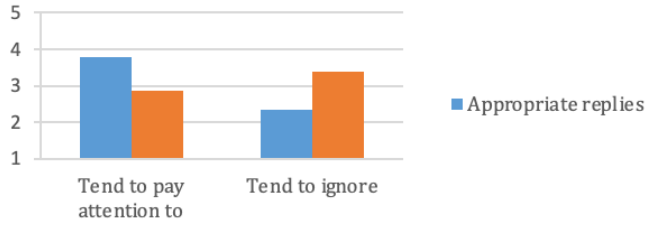
4.1 Attention/Ignore tendency for appropriate replies and inappropriate replies

The result from the attention/ignore tendency (Figure 3) showed that the participants tend to pay attention to the appropriate replies ($M = 3.78$, $SD = 1.07$) more than inappropriate replies ($M = 2.88$, $SD = 1.14$) ($t(97) = 4.521$, $p < .001$, $d = 0.457$) and ignore inappropriate replies ($M = 2.33$, $SD = 0.95$) more than appropriate replies

¹EasonC13, AI_Chatbot_experiment_backend: https://github.com/EasonC13/AI_Chatbot_experiment_backend

Table 3: The regression result for the frequency of suitable replies predicts CUQ score of the four groups. (Note: * $p < .001$).**

	Five-Replies-Anime		Five-Replies Anonymous		Single-Reply Anime		Single-Reply Anonymous	
The frequency of suitable replies	.576	7.046***	.514	5.997***	.339	3.607***	.360	3.860***
F	49.646***	35.969***		13.014***		14.897***		
Adjusted R2	.325	.257		.106		.121		

**Figure 3: Attention/Ignore tendency survey result for appropriate replies and inappropriate replies**

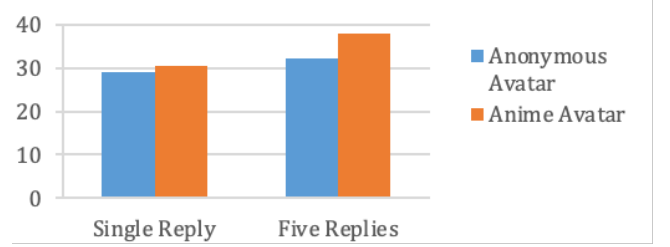
($M = 3.40$, $SD = 1.182$) ($t(97) = -5.890$, $p < .001$, $d = 0.595$). **H3 Accepted.**

4.2 The relation between the frequency of suitable replies and the CUQ Score

Four simple linear regressions were carried out to test if the frequency of suitable replies ($M = 3.88$, $SD = 1.58$) predicted the CUQ score of the four groups. All four results of the regressions are significant (Table 3). Indicated that the frequency of suitable replies significantly predicted chatting experience. Moreover, the result shows that the frequency of suitable replies can predict the CUQ score more on five-replies conditions. (H1 Accepted)

4.3 Comparison of the CUQ scores in the four groups

CUQ scores are subjected to a two-way analysis of variance for the following four conditions: 2 (single reply vs. five replies) \times 2 (anonymous avatar vs. anime avatar). The interaction effect for CUQ score between reply count and avatar type is non-significant ($F(1, 97) = 2.13$, $p = .148$). Therefore, H4 can be rejected. Regards to the CUQ scores for the four groups (Figure 4), from high to low are Five-Replies-Anime ($M = 37.5$, $SD = 20$), Five-Replies Anonymous ($M = 30.9$, $SD = 18.4$), Single-Reply Anime ($M = 30.5$, $SD = 18.8$), Single-Reply Anonymous ($M = 29.0$, $SD = 20.1$). Furthermore, the main effect of the number of replies indicates that for CUQ score, the five replies conditions are significantly higher than the single reply conditions ($F(1, 97) = 10.09$, $p = .002$, $\eta^2 = .094$). (H2 Accepted) Apart from that, the main effect of avatar type indicates that for CUQ score, the anime avatar conditions are significantly higher than the anonymous avatar conditions ($F(1, 97) = 4.57$, $p = .035$, $\eta^2 = .045$).

**Figure 4: The difference between the four groups.**

5 GENERAL DISCUSSION

The result indicates the following: 1. When users chat with an NLG Chatbot, they tend to pay attention to appropriate replies and ignore inappropriate replies. 2. When designing an NLG chatbot, if responding to a user's utterance with multiple replies, simulating a multi-person group chat will bring a better chatting experience. Therefore, it is easier to satisfy users with multiple responses, which can reduce the influence of occasional poor answers of the system. 3. There is no interaction effect between the avatar and the number of replies on the chatting experience. Thus, increasing the number of answers while providing multiple virtual avatars can bring a higher chatting experience score. Also, according to the effect size (η^2), to improve the chatting experience for an NLG chatbot which is single reply and anonymous avatar, providing multiple replies ($\eta^2 = .094$) might benefit more than setting an anime avatar ($\eta^2 = .045$). Our finding demonstrates the potential of letting chatbot to response to user's utterance with multiple replies. At present, most NLG chat-oriented chatbots only provide one reply. Therefore, we hope that developers can refer to this research and put multiple replies into design considerations when developing a chat-oriented chatbot in the future.

ACKNOWLEDGMENTS

This work was supported by the Ministry of Science and Technology of Taiwan (R.O.C.) under Grants 110-2813-C-003-033-E. We thank Yuen-Hsien Tseng, Tzung-Jin Lin, and Ching-Lin Wu for commenting on our manuscript and suggestion from Guo-Li Chiou, Chien Wen Tina Yuan, Tsung-Ren Huang and Liang-Yi Li.

REFERENCES

- [1] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. *Advances in Neural Information Processing Systems* 13 (2000).
- [2] John Brooke. 1996. Sus: a "quick and dirty" usability. *Usability evaluation in industry* 189, 3 (1996).
- [3] Lakindu Gunasekara and Kaneeka Vidanage. 2019. Unionbot: Semantic natural language generation based api approach for chatbot communication. In *2019*

- National Information Technology Conference (NITC)*. IEEE, 1–8. <https://doi.org/10.1109/NITC48475.2019.9114440>
- [4] Ryuichiro Higashinaka, Masahiro Mizukami, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, and Yuka Kobayashi. 2015. Fatal or not? Finding errors that lead to dialogue breakdowns in chat-oriented dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2243–2248.
 - [5] Samuel Holmes, Anne Moorhead, Raymond Bond, Huiru Zheng, Vivien Coates, and Michael McTea. 2019. Usability Testing of a Healthcare Chatbot: Can We Use Conventional Methods to Assess Conversational User Interfaces?. In *Proceedings of the 31st European Conference on Cognitive Ergonomics* (BELFAST, United Kingdom) (ECCE 2019). Association for Computing Machinery, New York, NY, USA, 207–214. <https://doi.org/10.1145/3335082.3335094>
 - [6] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751* (2020). arXiv:1904.09751 [cs.CL]
 - [7] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and usability engineering group*. Springer, 63–76. https://doi.org/10.1007/978-3-540-89350-9_6
 - [8] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.
 - [9] AM Rahman, Abdullah Al Mamun, and Alma Islam. 2017. Programming challenges of chatbot: Current and future prospective. In *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*. IEEE, 75–78. <https://doi.org/10.1109/R10-HTC.2017.8288910>
 - [10] KE Weick, KM Sutcliffe, and D Obstfeld. 2005. Organizing and the Process of Sensemaking. *Organisation Science*, 16 (4).
 - [11] Te-Lun Yang and Yuen-Hsien Tseng. 2020. Development and Evaluation of Emotional Conversation System Based on Automated Text Generation. *Journal of Educational Media & Library Sciences* 57, 3 (2020). [https://doi.org/10.6120/JoEMLS.202011_57\(3\).0048.RS.CM](https://doi.org/10.6120/JoEMLS.202011_57(3).0048.RS.CM)
 - [12] Yaoqin Zhang and Minlie Huang. 2019. Overview of the NTCIR-14 short text generation subtask: emotion generation challenge. In *Proceedings of the 14th NTCIR Conference*.